DOCUMENT RESUME

ED 360 840 FL 021 417

AUTHOR	Irvine, Aileen
TITLE	The Design and Validation of a Multi-Level Reading Comprehension Test.
REPORT NO	ISSN-0959-2253
PUB DATE	93
NOTE	8p.; For serial publication in which this paper appears, see FL 021 410.
PUB TYPE	Reports - Descriptive (141) -- Journal Articles (080)
JOURNAL CIT	Edinburgh Working Papers in Applied Linguistics; v4 p81-86 1993
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*English (Second Language); Foreign Countries; Higher Education; *Reading Comprehension; *Test Construction; *Test Validity
IDENTIFIERS	University of Edinburgh (Scotland)

ABSTRACT

Constructing a test of English-as-a-Foreign-Language reading comprehension that will accommodate the complete spectrum of performance from beginner to near native speaker can be problematic. Such a test is currently being developed at the Institute for Applied Language Studies at the University of Edinburgh (Scotland). With only a few hundred students to validate such a widely discriminating test, the problems become practical as well as theoretical. This article is a short report on why there was a need for the test and how the problems were approached. (Author/JL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 360 840

THE DESIGN AND VALIDATION OF A MULTI-LEVEL READING COMPREHENSION TEST

AILEEN IRVINE (IALS)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

BRIAN PARKINSON

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

FL021417

THE DESIGN AND VALIDATION OF A MULTI-LEVEL READING COMPREHENSION TEST

Aileen Irvine (IALS)

Abstract

Constructing a test of EFL reading comprehension which will accommodate the complete spectrum of performance from beginner to near native-speaker can be problematic. Such a test is currently being developed at the Institute for Applied Language Studies, University of Edinburgh. With only a few hundred students to validate such a widely discriminating test, the problems become practical as well as theoretical. This article is a short report on why there was a need for the test and how the problems were approached.

1. Background to the test

The Edinburgh Project on Extensive Reading (hereafter referred to as EPER) aims to offer a complete "extensive reading" package to its customers. That is to say that EPER not only organises the selection and dispatch to the client of appropriate EFL graded readers and accompanying back-up materials, but can provide tests for the placement of a student into a reading scheme and for the formal measurement of a student's progress within the scheme.

For future understanding of the rationale behind the test design, it is important to note here that EPER assigns each EFL graded reader to one of eight EPER levels, overriding the publishers' own level classification. The EPER levels are: G, F, E, D, C, B, A and X - G being the easiest and X the most difficult.

The purpose of a test is to decide which EPER level a student should "rightly" be reading at. From the user's point of view, this simply means that each possible raw score on the test should be interpretable in terms of an EPER reading level.

Until now, EPER has used a pair of standardized cloze tests to decide a student's initial reading level and, with subsequent administerings of the tests, to see how far up the EPER ladder of levels a student has moved. These tests have the advantages of being easy to administer and easy to mark, and the scores are immediately convertible to EPER reading levels using the scores conversion tables made available to EPER's customers.

However, there are also major disadvantages to these tests. One point of dissatisfaction is that, since the cloze tests are here being offered as tests of extensive reading competence rather than as tests of general language proficiency (which is the more widely accepted use of cloze tests), they do have extremely little face validity. Many teachers, students and administrators do not see any immediately visible connection between a cloze test score and extensive reading performance. The point about face

FL021417

validity is not whether the test is or is not an accurate measure of what it purports to measure, but whether it is seen to be such by the users. The cloze tests' lack of face validity in this context has proven to be unsettling for some of EPER's clients.

A second argument against the cloze tests is that they can be a very traumatic experience for lower level students. The two tests, Cloze A and Cloze B, are designed to measure all levels of proficiency. Although Cloze A is split into first half (the easier part) and second half (more difficult), Cloze B is given to students in its entirety. This means that the lower level students are faced with a test over half of which they have no hope of being able to complete. The weakest students may have to leave nine tenths of the answer-sheet blank. Unless Cloze A is to be given repeatedly, all the students in a reading programme which uses the tests will face Cloze B at some time in their extensive reading careers. In any case, Cloze A also, despite its division into two parts, will give rise to the same kind of situation, with some students effectively only able to perform on a quarter or a half of the test they have been given. This is not really acceptable for many teachers. It can also affect a test's reliability as a true measure of student performance, since a student who experiences a sense of defeat before even putting pen to answer-sheet will quite possibly perform a lot worse than he otherwise might have done.

These two concerns have provided a large part of the impetus for the development of the new extensive reading test.

2. Structure of the test

As it stands at the moment, the new test consists of two component parts - a reading comprehension paper and a separate discrete-item multi-choice vocabulary test.

2.1 Vocabulary test

Whereas the reading comprehension paper is stratified (and will be discussed later), the vocabulary test is a single unit and the same test is given to all students at every level. The vocabulary test consists of 70 items. Items are graded, so that the easier items are at the beginning, with the test becoming progressively more difficult.

The vocabulary test is subject to some of the same criticisms as the cloze tests. Firstly, with items designed to measure performance from beginner to near native-speaker level, it may be rather off-putting for the weaker students. Secondly, it has less face validity than the reading comprehension component which looks like a reading test, (although it has arguably more face validity than the cloze tests, since at least vocabulary acquisition is fairly easy to associate with extensive reading).

Two further arguments against the vocabulary test are found in the questions of construct and content validity. As Hatch and Farhady put it, the '...problem in construct validity is whether our test items really comprise the construct...' (Hatch and Farhady 1982: 252). In this particular case the problem would be whether a multi-choice vocabulary test can really be a measure of reading competence. Most of us would intuitively agree that an increase in a student's receptive vocabulary will be one of the gains from reading extensively, but most of us would also agree that the reading process must involve far more than receptive vocabulary knowledge. At best then, the vocabulary test can only comprise part of the extensive reading construct.

As for content validity - defined by Hatch and Farhady as ' ...the extent to which a test measures a representative sample of the subject matter content...' (Hatch and Farhady 1982: 251) - one problem might be simply that the vocabulary test is too short. Given that the test must discriminate eight different levels, and there are 70 items in the test, discrimination between any two adjacent levels will largely depend on a student's response to fewer than ten items. Are nine responses really enough to gauge a student's vocabulary knowledge? Even under the impossible circumstances of each item being perfectly reliable and each student and each marker performing perfectly reliably and there being no overlap between levels, it is not likely that nine vocabulary items could be properly representative of the vocabulary in the readers at any given EPER level.

To increase its content validity the vocabulary test would need to be made much longer - but then it would be even more like Cloze B, in that the longer the test, the greater the number of items which are inaccessible to the lower level students. One answer to this would of course be to split the test into two halves or stratify it even further into anything up to eight levels. The question of construct validity would however remain.

The present recommendation to the test's users in Hong Kong (which is where the test was piloted and which is, at the moment, the only place where it is used) is, where possible, to use scores from the reading comprehension component rather than from the vocabulary test component to determine a student's EPER reading level. It is even possible that the final version of the extensive reading test will be reduced to the reading comprehension component only. Quite apart from the other considerations, the more compact a test, the more administratively practical it is - particularly important where thousands of testees are to be involved. Whether the vocabulary test will eventually be removed, or whether it will remain in some kind of complementary capacity to the reading comprehension component will depend on the results of the follow-up research into the test's use in Hong Kong which is planned for 1993.

2.2 Reading comprehension papers

Whereas the vocabulary test is a single unit intended to accommodate all levels, the reading comprehension paper is in fact eight separate reading comprehension papers - one for each EPER reading level. The only feasible way to have one single reading comprehension paper for all levels would be to have one paper made up of increasingly difficult comprehension passages. Honouring the distinction between "extensive" and "intensive" reading as best we can (extremely difficult when attempting to write a test since testing is by nature an intensive activity, unlike continuous assessment), and having rather longer than usual passages, this could have resulted in a three- to four-hour reading comprehension paper. Again, the weaker students would be demoralised and class time would be wasted as they sat for three hours in front of an impossible task. Conversely, the stronger students would be wasting their time on material far too easy for them, and marking would take longer. Moreover, the students would end up sitting the whole composite paper again and again. In view of all this, it seemed better to have eight separate papers.

The eight papers were then grouped as four pairs of adjacent papers. Each student takes a pair of papers - the combined scores on two papers giving a more reliable result than a score from one paper only. The decision as to which pair of papers a student will take ultimately rests with the teacher and will depend upon the student's current reading level. However, if the student gets above a certain combined score on a pair of reading comprehension papers, then the student should be given the two reading comprehension

papers immediately above. Likewise if a student obtains below a certain score then he should be given two easier reading comprehension papers. The middle range of scores from any two paired papers is divided into two bands, each band of scores pertaining to one EPER reading level. Thus a student need not take more than four papers (two pairs), but the vast majority of students will need to take only one pair of papers. This will not only save everyone's time, but will cut down on student frustration arising from being faced with material which is far too easy or far too difficult.

So far as we know, the Hong Kong administration is very pleased with the new more user-friendly reading comprehension test, which also has more face validity than the old cloze tests. It is also my personal belief that teachers in Hong Kong will feel more personally involved in the test than they did with the old cloze tests, since their initial judgement on a student's EPER reading level is what is used to route the student towards the appropriate pair of tests. Teachers are thus asked to make a professional contribution to the testing machinery, something they were not asked to do with the cloze tests where all students automatically took the same test. I believe that this professional involvement will have a favourable effect on teachers' attitudes towards the test. (Whether this is indeed the case will be researched in the 1993 follow-up study.)

3 Test validation

Both the vocabulary test and the reading comprehension papers were piloted in Hong Kong, and the results analyzed at Edinburgh.

3.1 Validation of the vocabulary test

Analysis of the vocabulary test was very straightforward. A Rasch analysis gave each item a difficulty estimate and these difficulty estimates were then used to convert each possible raw score on the vocabulary test to a student ability estimate. A scale of abilities was then devised with eight student ability bands - each band corresponding to one EPER reading level. The top and bottom cut-off points for each band (and hence for each EPER reading level) were decided through post-hoc comparison with known cloze scores (each student in the pilot already having been assigned an EPER reading level on the basis of a cloze score). That is to say that, for example, the typical ability estimates for students already assigned to EPER reading level B through their cloze scores would be used as the student ability estimates for band B on the vocabulary test and students demonstrating those same abilities on the vocabulary test would be assigned to level B. In other words, the vocabulary test was validated and stratified against the cloze tests.

This obviously raises the question of whether such a validation is really tenable, given that a cloze test and a multi-choice vocabulary test are two quite different testing beasts and may be measuring quite different things. The correlation between the cloze scores and the vocabulary test scores was however .8 and it will be interesting to see how well the comparison between the two test-types holds in practice.

3.2 Validation of the reading comprehension papers

Validation of the reading comprehension component was a little less straightforward. Although the eight reading comprehension papers were conceived in such a way that

putting them all together would provide a complete test of reading comprehension ability from beginners to very advanced, no student sat the complete test. Thus a correlation between reading comprehension scores and existing cloze scores or the vocabulary test scores would have been nonsensical. (For example a student sitting the two lowest level reading comprehension papers may well have obtained a very high score on these, but - being at a lower level - would have obtained a very low cloze score. A student who obtained a very high score on the cloze, but who sat the highest level reading comprehension papers, might have a lower reading comprehension score than the low level student.)

Eight separate correlations - i.e. between the scores for each separate level of the reading comprehension test and cloze scores - might have made more sense. However there was not a wide enough range of scores at certain levels of the reading comprehension papers for a satisfactory correlation to be made. Nor was there a large enough number of scores at certain levels. Although the total number of students taking part in the pilot was over 200, the number for each level was sometimes less than thirty. (Hatch and Lazaraton (1991: 550) recommend a minimum N of 35.)

The other problem with the series of reading comprehension papers was how to interpret scores across papers. To give an example: if a student obtained a score of 30 on the lowest level pair of papers, then six months later obtained a score of 20 on the immediately higher pair of papers, how would a teacher know how much progress, if any, had been made by that student, given that the papers are pitched at different levels?

The obvious answer is regression, which would compute equivalent scores on different pairs of papers, but, as with correlation, the scores on the reading comprehension papers did not meet the technical requirements for regression. (Notably the numbers at certain levels were again too low, and the ranges of scores at some levels were not wide enough to establish a correlation - a high correlation between two groups of scores being a pre-requirement of regression.)

The procedure in fact followed was to place all eight reading comprehension papers consecutively one after the other to form one long test and to perform a Rasch analysis on all the items at the same time. This was possible because, although no student took more than two adjacent levels of reading comprehension papers, there was linking throughout the complete series of papers. That is to say that, for example, some students took the levels G and F papers and others took the levels F and E papers. The results of the F papers would provide the link between the G and E papers necessary to standardize item difficulty estimates across the three levels on to a common scale. With linking throughout the eight levels, item difficulty estimates would be standardized to a common scale throughout the whole test.

One problem with this, however, is that a reading comprehension test is not an ideal candidate for Rasch analysis, since the Rasch model assumes that all items are discrete, and that the chances of an item being answered correctly or incorrectly are unaffected by answers on preceding items. This is clearly not always the case in a reading comprehension test, although it does depend on the particular items and attempts were made at the item-writing stage to produce independent items. Again, it will be interesting to find out how well the Rasch item difficulty estimates for the reading comprehension papers perform in practice.

The Rasch item difficulty estimates were then used to produce student ability estimates from raw scores on pairs of papers. To give an example: a combined score of 30 on the levels F and G papers would show an ability rating of -0.2, which is the same ability rating as a combined score of 18 on the levels D and E papers.

The next logical step would be a scores conversion chart. However there are many factors affecting the reliability of such ability estimates, and in order to avoid any spurious accuracy the scores were not reported as individual scores and equivalent scores on different pairs of papers, but again a bands system was used. The same procedure was used to decide upper and lower ability estimates for each band as was used for the vocabulary test bands. The ability estimates of students known to be reading at a certain level (allocated to that level because of their cloze scores) were taken as the prescriptive ability estimates for that EPER reading level.

Thus the reading comprehension component, like the vocabulary component, was validated against the cloze tests. The same question arises - how legitimate was this procedure? The follow-up study will hopefully give some indication.

4. Conclusion

One of the main problems in constructing a test for all levels is the avoidance of material which is much too easy or much too difficult for some of the students. The answer to this in our case was a stratified test divided into eight levels. However, this stratification brought with it a number of problems at the validation stage, particularly how to standardize scores at different levels on to a common scale. Several standard statistical procedures (correlation and regression) could not be used because of the low numbers of students who took part in the pilot at each level. Our answer was to treat the eight papers as one continuous test and then perform a Rasch analysis. We then used Rasch item difficulty estimates and student ability estimates to compute equivalent scores across levels.

An extensive follow-up study for the test is planned for 1993. How well this validation procedure works in practice will then be investigated.

References

Hatch E. and H. Farhady. 1982. Research Design and Statistics for Applied Linguistics. Rowley, Massachusetts: Newbury House.

Hatch E. and A. Lazaraton. 1991. The Research Manual: Design and Statistics for Applied Linguistics. Rowley, Massachusetts: Newbury House.